The Second International Symposium on Optimization and Systems Biology (OSB'08)
Lijiang, China, October 31– November 3, 2008

Distinguishing enzymes from non-enzymes via support vector machine

Yongcui Wang[1] Yingjie Tian[2] Naiyang Deng[1,*]

[1]College of Science, China Agricultural University, Beijing, China, 100083
[2]Research Center on Fictitious Economy & Data Science Chinese Academy of Sciences, Beijing, China, 100190

Abstract With many proteins sequenced, the ability of predicting protein function from sequence is becoming more and more important. Currently, methods for inference of the protein functional annotation are mostly based on identifying a known function protein which is similar to the query protein. However, for the proteins that are dissimilar or only similar to the unknown proteins, these methods will lose effectiveness. In this paper, we propose a new method for distinguishing enzymes from non-enzymes without similarity search. We use conjoint triad feature, secondary-structure content and surface pocket properties to describe 1178 high-resolution proteins, and apply support vector machine approach to assign these described proteins class. With 10-fold cross-validation, the accuracy of predicting functional class of enzymes and non-enzymes is about 85.19%. Moreover, by choosing the 'informative' features, the accuracy can be improved to 86.31%. These results suggest that this newly sequence-based method can be used to discover the other functional class membership of proteins.

Keywords Protein function; Functional class of enzymes and non-enzymes; Support vector classification; Feature vectors;

1 Introduction

With many proteins sequenced, the ability of predicting protein function from sequence is becoming more and more important. Among all groups of proteins, enzymes are the largest and most diverse one and they catalyze all chemical reactions in the metabolism of all organisms[18], so the ability of predicting functions of enzymes is essential for understanding molecular mechanisms. However, inference of functional annotation for a newly-sequenced protein is still a time-consuming and costly task purely relying on biochemical experiments. It is highly desired to develop an efficient computational method to implement this task.

As one kind of computational method, similarity search among proteins with known function is the basis of current function prediction methods [16, 11]. They deduce the protein function by identifying sequence or structural similarity to a known function protein. However, for the proteins that are dissimilar or only similar to the unknown proteins,

*Corresponding author. e-mail: dengnaiyang@vip.163.com

these methods will lose effectiveness. In order to remedy this case, many machine learning methods were used, which obtained encouraging results. For example, as an excellent machine learning method, support vector machines (SVMs) which are motivated by statistical learning theory [20, 21], have been successfully applied to a wide range problems in bioinformatics, including cancer classification [3, 17, 12], protein-protein interaction prediction [1], protein fold recognition [8], protein structure prediction [6] and so on. They also serve as effective methods for protein function classification [5, 9]. In [9], the authors described proteins using features of secondary-structure content, amino acid propensities, surface properties and ligands, and they applied support vector classification (SVC) to predict enzyme function, the accuracy of which is about 80.17% . In [2], the authors transformed the sequence, second structure, surface and biochemical information of a protein to a graph, by using graph kernel and SVC on these protein graphs, they obtained an improved predicting accuracy of 84.04% on the same dataset (created by Dobson and Doig [9]).

A lot of approaches of feature representation have been used in prediction of protein function, including amino acids composition (AAC). The AAC feature representation is denoted by a 20-dimensional vector which consists of occurrence frequencies of the amino acids. Owing to lack of the sequence order information, some modified version of AAC such as pseudo amino acid composition (Pse-AAC, Chou and Elrod, 2003) and amphiphilic pseudo-amino acid composition (Am-Pse-AAC, Chou, 2005) have been developed. However, both Pse-AAC and Am-Pse-AAC have some parameters to be determined, and also need the property of physio-chemistry of amino acids. Therefor, it is highly desired to develop a simple and efficient feature representation method.

Recently, Shen et al. propose a simple method based only on the information of protein sequences for protein-protein interactions (PPIs) prediction [19]. They used SVC with a conjoint triad feature for describing proteins, obtained much higher prediction accuracy than other sequence-based PPIs prediction methods. Inspired by this, we present a new method that combines secondary-structure content and the size of largest surface pocket with this conjoint triad feature to describe proteins. We also apply SVC to predict enzyme function and test its performance on the dataset used in [9].

2 Materials and methods

2.1 Material

The dataset used here is same to that used in [2, 9], which was obtained from PDB database. To reduce the homology bias, a redundancy cutoff was operated such that no chain in any protein aligns to any other chain in the same functional class with a Z-score$\geq$ 3.5. Finally, 691 enzymes and 487 non-enzymes were obtained. The detailed description can be found in [9].

2.2 Method

2.2.1 Support vector classification

Since SVMs were proposed in the 1990s, they have been successfully applied to a wide range of pattern recognition problems including handwriting recognition, object recognition, face detection, text categorization and so on [4]. Now we briefly introduce the standard algorithm *C*-SVC for classification problems:

For the given training set

$$T = \{(x_1, y_1), \cdots, (x_l, y_l)\} \tag{1}$$

with input $x_i = (x_{i1}, \cdots, x_{in})^T \in \mathrm{R}^n$ and output $y_i \in \{-1, +1\}$, where x_{ij} represent the jth feature of the ith feature vector $\boldsymbol{x}_i$. Let $\phi : R^n \to \mathscr{H}$ be a mapping from input (feature) space to a Hilbert space $\mathscr{H}$. C-SVC finds a hyperplane $(w \cdot \phi(x)) + b = 0$ which can separate the two classes with the maximal margin and minimal training errors in the Hilbert space. By applying a kernel function to replace the inner product in $\mathscr{H}$, the corresponding decision function is

$$f(\boldsymbol{x}) = \mathrm{sgn}(\sum_{i=1}^{l} \alpha_i^* y_i K(\boldsymbol{x}_i, \boldsymbol{x}) + b^*), \tag{2}$$

where α^* is the solution of the following optimization problem

$$\min_{\alpha} \quad \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} y_i y_j \alpha_i \alpha_j K(\boldsymbol{x}_i, \boldsymbol{x}_j) - \sum_{j=1}^{l} \alpha_j, \tag{3}$$

$$\text{s.t.} \quad \sum_{i=1}^{l} y_i \alpha_i = 0, 0 \leqslant \alpha_i \leqslant C,\ i = 1, \cdots, l, \tag{4}$$

and b^* can be obtained as follows: if there exist $\alpha_j^* \in (0, C),\ j = 1, \cdots, l$, then

$$b^* = y_j - \sum_{i=1}^{l} y_i \alpha_i^* K(\boldsymbol{x}_i, \boldsymbol{x}_j). \tag{5}$$

Many kernel functions can be used in C-SVC, including polynomial kernel $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = (1 + (\boldsymbol{x}_i \cdot \boldsymbol{x}_j))^d$, RBF kernel $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\left(-\frac{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|}{2\sigma^2}\right)$ and so on. In [9], RBF kernel is used because it perform well than other kernels, so in this paper, we use RBF kernel function to test the performance of our method.

2.2.2 Construction of feature vectors

1. Conjoint triad(CT) feature vectors

Construction of feature vectors for each type of data dominates the learning capability of the SVC. Since applying the conjoint triad feature with SVC can obtain the promising results for predicting PPIs. This imply the efficiency of such description, so we follow the idea here. Now we briefly introduce this feature vector construction method.

Classifying amino acids. Based on the dipoles and volumes of the side chains, the 20 amino acids can be classified into seven classes. Amino acids within the same class likely involve synonymous mutations because of their similar characteristics.

Constructing feature vectors by conjoint triad [19]. Each protein sequence can be projected into a vector by counting the frequencies of each amino acid triad (any three continuous amino acids) type.

First, representing each protein sequence by a binary space (V, F), where V=$(v_1, \cdots, v_m)$ represents the vector space of sequence features, and each feature v_i represents a sort of triad type; $F = (f_1, \cdots, f_m)$ is the frequency vector corresponding to V, and the value of

the f_i is the frequency of type v_i appearing in the corresponding protein sequence. Because the amino acids have been catalogued into seven classes, the size of V should be $7\times 7\times 7 = 343$; thus, $m = 343$. The detailed definition for (V,F) are illustrated in [19]. Clearly, the value of f_i correlates to the length (number of amino acids) of a protein. In general, a long protein would have a large value of f_i, which complicates the comparison between two heterogeneous proteins. To solve this problem, d_i as normalized f_i is defined with the following equation:

$$d_i = \frac{f_i - \min\{f_1, f_2, \cdots, f_{343}\}}{\max\{f_1, f_2, \cdots, f_{343}\}}. \tag{6}$$

Thus, a 343-dimension vector corresponding to a protein can be constructed.

2. Secondary structural(SS) features

For secondary structural property which can be obtained by the the algorithm of GOR [22] , features are extracted from primary sequence based on three descriptors: 'composition': percent composition of 3 constituents (e.g. α-helic, β-sheet, coil); 'transition' : the transition frequencies(α-helic to β-sheet, β-sheet to coil etc.); and 'distribution': the distribution pattern of constituents (where the first residue of a given constituent is locates, and where 25%, 50%,75% and 100% of that constituent are contained). For concrete details, see [10]. Then, a 21-dimension vector corresponding to a protein can be constructed.

3. Surface features

For surface property which can be calculated by the cavity detection algorithm CASTp [23] , two features are extracted separately from the area and volume of largest surface pocket, which represent the size of surface pocket for a protein. Then, a 2-dimension vector corresponding to a protein can be constructed.

4. 'Informative' features

From these three above type features (343+21+2), the 'informative' features were selected via the following criteria (F-score): first, for the jth feature, calculate

$$P(j) = \frac{(\mu_1(j) - \mu(j))^2 + (\mu_{-1}(j) - \mu(j))^2}{\sigma_1^2(j) + \sigma_{-1}^2(j)},\ j = 1, \cdots, n,$$

where $\mu_1(j)$ and $\mu_{-1}(j)$ are the mean values of jth feature of all inputs in class $+1$(enzyme) and class -1 (non-enzyme) respectively, where $\mu(j)$ is the mean value of jth feature of all input; $\sigma_1(j)$ and $\sigma_{-1}(j)$ are the standard deviations of jth feature of all inputs in class $+1$ and class -1 respectively. Then rank $P(j)\,(j = 1, \cdots, n)$ in descending order and choose the top 120 corresponding features as the 'informative' features. Note that, the number of the 'informative' features (120) is determined by the cross-validation.

3 Results and discussion

In this section, we evaluate the performance of the new method for distinguishing enzymes from non-enzymes. The experiments were implemented by Libsvm (version 2.84) [15].

3.1 Parameters selection

The performance of C-SVC heavily depends on the selection of several parameters. In our experiments, parameters C, σ should be appropriately chosen, as C controls the tradeoff between maximizing the margin and minimizing the training error, and σ dominate the generalization ability of SVC by regulating the amplitude of the RBF kernel function. For every test, we selected them by 3-fold cross-validation in the training set, and C was chosen from the set $\{0.25, 0.5, 1, 5, 10, 100, 1000\}$, σ from the set $\{0.1, 0.25, 0.5, 1, 5, 10, 100\}$, then the optimal parameter pairs (C^*, σ^*) were used in SVC.

3.2 Comparison with other methods

We constructed 5 different datasets by using different feature vectors, such as the dataset with only conjoint triad features (i.e. the data with 343 dimensions), the dataset with conjoint triad features and secondary structure features (343+21 dimensions), and so on. Then for each dataset we perform the C-SVC algorithm.

The results of average prediction accuracy and the standard deviation in 10-fold cross-validation with C-SVC are listed in table 1. For example, using the conjoint triad feature vector (343 dimensions) in C-SVC, the average accuracy is 82.48% , it is better than the method [9] which requires additional information (such as ligands, surface pocket, secondary structure and bonds of the protein).

We can see in Table 1 that the average classification accuracy corresponding to 366-dimension vector is 85.19%, which is not only better than the vector model proposed by Dobson and Doig and graph kernel model proposed by Smola, but also better than the result of conjoint triad feature vector (343-dimension). This suggests that the information of secondary structure and surface can improve the prediction accuracy, and are important to determine the functional class of enzyme. Furthermore, by selecting the 120 'informative' features in 366 features, the accuracy was improved to 86.31%. Among these 'informative' features, 104 are sequence features, 14 are secondary structure features and 2 are surface features. Surface of protein is further proved to be important to distinguish enzymes from non-enzymes, and this result also imply that there exist some relative small regions in protein sequence which determine the functional class membership of protein. Future work can focus on detecting these important sequence regions to improve the accuracy of prediction.

3.3 The other evaluation criterion

Moreover, enzymes is the most focus of our attention, since it is important for further study, including predicting the family and subfamily of enzymes and so on, so precision($\frac{true\ positive}{true\ positive + false\ positive}$) and sensitivity ($\frac{true\ positive}{true\ positive + false\ negative}$) are used to evaluate the the ability of identifying enzymes of our method. Because the results of Dobson and Smola methods under these criterion were not reported on the corresponding literatures [2, 9], and we didn't construct the feature vector as Dobson and Smola done for simplicity, so we only reported the results of our method under these criterion in Table 2. The precision is in the range of $85\% - 87\%$, the sensitivity is in the range of $87\% - 89\%$. The good performance of present method can be seen clearly.

Table 1: The first line list the results obtained in[9], the second line list the results obtained in[2], the following lines lists the results of the conjoint triad feature vector construction method, the conjoint triad method add up the information of secondary structure, surface pocket, secondary structure and surface pocket respectively, the last line is the results obtained from 'informative' features-based method.

Feature type	*Accuracy%*	*St.dev*
Features in [9]	80.17	1.24
Graph fature in [2]	84.04	3.33
CT (343)	82.48	2.46
CT (343) + *SS* (21)	83.62	2.84
CT (343) + *surface* (2)	84.78	1.69
CT (343) + *SS* (21) + *surface* (2)	85.19	2.03
'*Informative*' *features* (120)	86.31	1.19

Table 2: The performance of the new methods

Feature type	*Sensitivity%*	*Precision%*
CT (343)	87.32 ± 3.59	85.14 ± 2.69
CT (343) + *SS* (21)	88.38 ± 2.64	86.75 ± 2.77
CT (343) + *surface* (2)	88.94 ± 1.43	84.74 ± 2.37
CT (343) + *SS* (21) + *surface* (2)	89.89 ± 1.86	85.91 ± 1.57
'*Informative*' *features* (120)	89.55 ± 1.59	87.67 ± 1.88

Table 3: The results of predicting 56 no-functional annotations proteins

Accuracy%	*Sensitivity%*	*Precision%*
88.46	88.89	88.89

3.4 Results of 56 no-functional annotations proteins

There are 56 protein which have no-functional annotations in the dataset provided by Dobson and Doig, now these functional class can be obtained by PDB and ExPASy database. We use the present method to identify whether they are enzymes or not, and the results are listed in table 3. Among 27 enzymes in these 56 proteins, we only missed 3 enzymes.

4 Conclusions

In this paper, a new method for predicting enzyme function was proposed. The present method combines structure and surface information with a conjoint triad feature to describe proteins and applied SVC on these described proteins. Even only use the conjoint triad feature vector in SVC, the average accuracy is better than the method in [9] which requires additional information (such as ligands, surface pocket, secondary structure and bonds of the protein). This imply the rich information in sequence is conclude in conjoint triad feature. After adding structure and surface information into these conjoint triad feature vectors, more better results can be obtained. Furthermore, by choosing the 'informative' features , more promising results can be obtained. These results suggest that our new method for predicting enzyme function can also be used to identify the other protein

functional class membership.

In this paper, we focus on the proteins whose surface pockets are exist. Once the target protein can't find any pocket, one can use zero to replace the corresponding feature value. Alternatively, this missing feature can also be considered as structural absent, you can use the modified version of SVM [7] to solve this problem.

From the present work, we feel that using the more relevant feature to class membership, the higher accuracy of prediction will be obtained. So future work will aim at finding the highly relevant features to the functional class membership and design the efficient feature vector construction method to improve the prediction accuracy.

Acknowledgments

This work is supported by the Key Project of the National Natural Science Foundation of China(No. 10631070), the National Natural Science Foundation of China(No. 10601064, No. 70601033).

References

[1] Bock, J.R.D., Gough, A., 2001. Predicting protein–protein interactions from primary structure. *Bioinformatics* **17**, 455–460.

[2] Borgwardt, K.M., Ong, C.S., Vishwanathan, S.V.N., Smola, A.J., Kriegel, H., 2005. Protein function prediction via graph kernels. *Bioinformatics* **21**, i47–i56.

[3] Brown, M.P.S., Grundy, W.N., Lin, D., Christianini, N., Sugnet, C., Ares, M., Haussler, D., 1999. Support vector machine classification of microarray gene expression data, UCSC-CRL 99-09 . Department of Computer Science, University California Santa Cruz. Santa Cruz. CA.

[4] Burges, C. J.C. ,1998. A tutorial on support vector machines for pattern recognition. *Data Ming and knowledge Discovery* **2(2)**, 121–167.

[5] Cai, C.Z., Wang, W.L., Sun, L.Z., Chen, Y.Z., 2003. Protein function classification via support vector machine approach. *Mathematical Biosciences* **185(2)**, 112–122.

[6] Cai, Y.D., Liu, X.J., Xu, X.B., Chou, K.C., 2002. Prediction protein structural classes by support vector machines. *Computational Biology and Chemistry* **26**, 293–296.

[7] Chechik, G., Heitz, G., Elidan, G., Abbeel, P., Koller, D., 2008. Max-margin classification of data with absent fearture. *Journal of Machine Learning Research* **9**, 1–21.

[8] Ding, C.H. Q., Dubchak, I., 2001. Multi-class protein fold recongnition using support vector machines and neural network. *Bioinformatics* **17**, 349–368.

[9] Dobson, P.D., Doig, A.J., 2003. Distinguishing enzyme structures from non-enzymes without alignments. *Journal of Molecular Biology* **330(4)**, 771–783.

[10] Dubchak, I., Muchnik,I., Holbrook, S.R., Kim, S.H., 1995. Prediction of protein folding class using global description of amino acid sequence. *Proceedings of the National Academy of Sciences* **92**, 8700–8704.

[11] Fetrow, J.S., Skolnick, J., 1998. Method for prediction of protein function from sequence using the sequence-to-structureto- function paradigm with application to glutaredoxins/thioredoxins and T-1 ribonucleases. *Journal of Molecular Biology* **281**, 949–968.

[12] Furey, T.S., Cristianini, N., Duffy, N., Bednarski, D.W., Schummer, M., Haussler, D., 2000. Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics* **16**, 906–914.

[13] Golub, T., Slonim, D., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D., Lander, E.S., 1999. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science* **286**, 531–537.

[14] Hegyi, H., Gerstein, M., 1999. The relationnship between protein structure and function: a comprehensie survey with application to the yeast genome. *Journal of Molecular Biology* **288(1)**, 147–164.

[15] Hsu, C.W., Chang, C.C., Lin, C.J., 2007. A practical guide to Support Vector Classfication. Available from: *http://www.csie.ntu.edu.tw/ cjlin*.

[16] Koonin, E. V., Tatusov, R. L., Galperin, M. Y., 1998. Beyond complete genomes: from sequence to structure and function. *Current Opinion in Structural Biology* **8**, 355–363.

[17] Mukherjee, S., Tamayo, P., Slonim, D., Verri, A., Golub, T., Mesirov, J.P., Poggio, T., 2000. Support vector machine classification of microarray data, AI memo 182. CBCL paper 182. MIT. Can be retrieved from ftp://publications.ai.mit.edu.

[18] Schomburg, I., Chang, A., Ebeling, C., Gremse, M., Heldt, C., Huhn, G., Schomburg, D., 2004. BRENDA, the enzyme database: updates and major new developments. *Nucleic Acids Research* **32**, D431–D433

[19] Shen, J.W., Zhang, J., Luo, X.M., Zhu, W.L., Yu, K.Q., Chen, K.X., Li, Y.X., Jiang, H.L., 2007. Predicting protein-protein interactions based only on sequences information. *Proceedings of the National Academy of Sciences* **104**, 4337–4341.

[20] Vapnik, V., 1995. The Nature of Statistical Learning Theory. Springer.

[21] Vapnik, V., 1998. Statistical Learning Theory. Wiley.

[22] http://npsa-pbil.ibcp.fr/cgi-bin/nps$a_a$utomat.pl?page=nps$a_g$or 4.html

[23] http://sts.bioengr.uic.edu/castp/